

# 本周周报

解聪

2013.12.2-2013.12.8

## 本周工作

对先前的工作进行了进一步的总结。我们先前做的可以认为是对离散变量的条件概率可视化。离散随机变量就是之前说的类别型数据。

维度可以进一步可以细分为行为的离散变量  $X$ （比如购买商品的类目），和反应用户基本属性的离散变量  $Y$ （比如用户的年龄性别等）。

### 1. 边缘概率，联合概率，以及条件概率的可视化：

以淘宝数据为例， $X$  表示商品购买类目， $Y1$  表示年龄， $Y2$  性别， $Y3$  星座。

我们将  $P(X|Y1,Y2,Y3)$  这样的条件概率在每一个方块中展示。

边缘分布也可以通过图 1 中的直方图展示出来，如  $P(Y3)$  和  $P(Y1,Y2)$ 。

每个方块上的谁水平条形图展示的是联合分布  $P(Y1,Y2,Y3)$

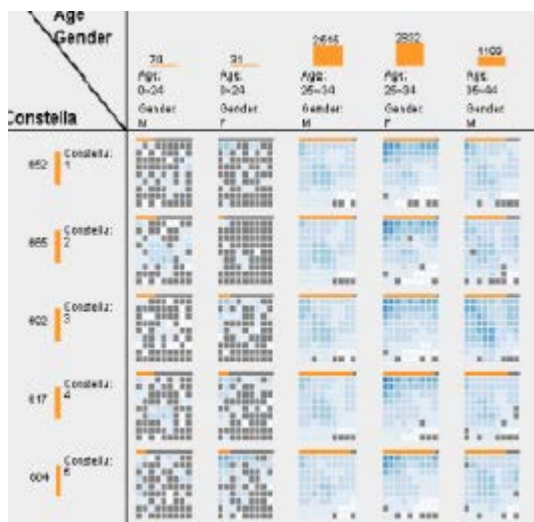


图 1

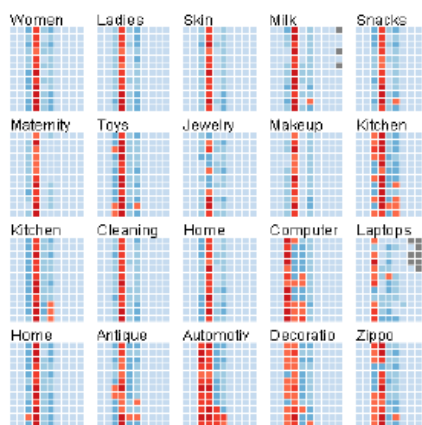


图 2

同时，我们也可以展示后验概率。如图二展示  $P(Y1,Y2,Y3|X)$  的条件概率：

### 2. 对条件概率分布的相似性的定义与用户交互

#### 1. 离散变量条件概率的聚类。可以采用以下几种方法：

矩阵的聚类：将每一行，或每一列做为基本单位进行聚类，优点是保留了边缘分布  $P_{\text{margin}}$ ，更易于发现随机变量不同取值之间的关系。具体做法是矩阵的行列重排序。

区块的聚类：直接不考虑边界分布，直接把每个条件概率  $P_{\text{Condition}}$  当做聚类的对象进行分析，优点是更容易发现单个条件概率之间的关系。具体做法是投影与普通聚类算法。

#### 2. 离散变量条件概率的相似查找

在第一步的基础上，用户如果想要寻找某一个选定的条件概率  $P_{\text{Condition}}$  的相似的条件概率分布。

可以先进行也可以局部可视化与矩阵的筛选，再通过人眼观察或者进行自动遍历寻找。

如果用户只需要局部的相似性，用户对矩阵的选取，像素块的选取来完成局部相似的寻找。

#### 3. 离散变量条件概率的相似度调整

这在之前周报里有提到，具体做法是利用用户对矩阵的选取，像素块的选取干预，修改距离矩阵的定义。

### 3. 对联合概率维度的选取

由于不同维度本身的熵不一样，维度之间熵也存在冗余，所以可以结合用户的交互以及需求来选取所需要的维度。

1. 离散变量本身的熵可视化
2. 离散变量的相对熵的可视化
3. 用户交互的选取与修改离散变量的关系

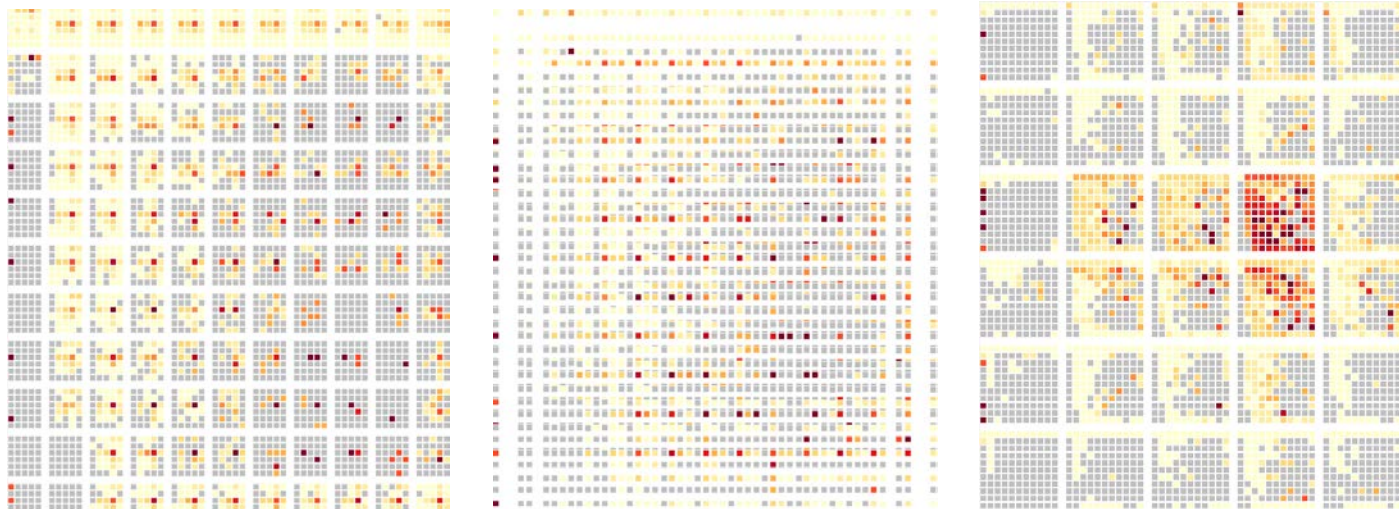
#### 4.先验概率与后验概率关系的可视化与贝叶斯公式：

$$P(X_1, X_2 \dots X_m | Y_1, Y_2 \dots, Y_n) = P(Y_1, Y_2 \dots, Y_n | X_1, X_2 \dots X_m) * P(X_1, X_2 \dots X_m) / P(Y_1, Y_2 \dots, Y_n)$$

这也是朴素贝叶斯分类器的理论基础，从语言上表述可以理解为  $\text{posterior} = \text{prior} * \text{likelihood} / \text{evidence}$ ;

所以需要展示先验概率和后验概率的对应关系。

我本周做了一个基本的过渡动画效果，下图从左到右展示了先验概率  $P(X \text{ 地点}, X \text{ 行业} | Y \text{ 广告}, Y \text{ 利润})$ ，动画的转换，以及  $P(Y \text{ 广告}, Y \text{ 利润} | X \text{ 地点}, X \text{ 行业})$ 。最右图的颜色映射还是采用的最左图的方案，即在最左图中按照每个方块内最大和最小值归一化数值。



#### 5. 其他

1. 由于数据中很多维度是连续型随机变量，所以需要将数值型变量转化为离散型变量。  
在对连续数值维度的划分过程中可以结合用户调整
2. 由于我们使用了图像的方式来展示了离散变量的概率分布，可能会使用到模糊集的相关方法。

本周时间主要在实现问题 3, 4 的想法，暂时还没有进一步的效果。其他的时间在写毕业论文和改专利。

#### 下周工作：

实现问题 3 和 4 中未完成的部分。

研究朴素贝叶斯分类器与本可视化方法结合的一些想法。